This research was funded in whole, or in part, by the Wellcome Trust [093820/Z/19/Z]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Testing and Correcting for Weak and Pleiotropic Instruments in Two-Sample Multivariable Mendelian Randomisation

Eleanor Sanderson[1,2], Wes Spiller[1,2] and Jack Bowden[3,1]

1. MRC Integrative Epidemiology Unit, University of Bristol.
2. Population Health Sciences, University of Bristol.
3. College of Medicine and Health, University of Exeter.

June 2021

Abstract

Multivariable Mendelian Randomisation (MVMR) is a form of instrumental variable analysis which estimates the direct effect of multiple exposures on an outcome using genetic variants as instruments. Mendelian Randomisation and MVMR are frequently conducted using two-sample summary data where the association of the genetic variants with the exposures and outcome are obtained from separate samples. If the genetic variants are only weakly associated with the exposures either individually or conditionally, given the other exposures in the model, then standard inverse variance weighting will yield biased estimates for the effect of each exposure. Here we develop a two-sample conditional F-statistic to test whether the genetic variants strongly predict each exposure conditional on the other exposures included in a MVMR model. We show formally that this test is equivalent to the individual level data conditional F-statistic, indicating that conventional rule-of-thumb critical values of $F > 10$, can be used to test for weak instruments. We then demonstrate how reliable estimates of the causal effect of each exposure on the outcome can be obtained in the presence of weak instruments and pleiotropy, by re-purposing a commonly used heterogeneity Q-statistic as an estimating equation. Furthermore, the minimised value of this Q-statistic yields an exact test for heterogeneity due to pleiotropy. We illustrate our methods with an application to estimate the causal effect of blood lipid fractions on age related macular degeneration.

1 Introduction

Instrumental variables (IV) is a form of regression analysis which estimates the causal effect of an exposure on an outcome in the presence of unobserved confounding. Mendelian randomisation (MR) is a rapidly expanding application of the IV method in the field of epidemiology in which genetic variants are used as instruments. If genetic variants - usually single nucleotide polymorphisms (SNPs) - are available which reliably predict the exposure and are not associated with the outcome through any other pathway, then they are valid IVs. These genetic variants can then be used as instruments to obtain an estimate for the causal effect of a modifiable health exposure on a disease outcome[1,2]. The results of such an analysis can inform the development of public health, or even pharmaceutical, interventions. MR is often conducted with summary -level data on the SNP-exposure and SNP-outcome associations obtained from genome-wide association studies (GWAS) without the need to have individual level data on the genetic variants, exposure and outcome available to the researcher conducting the MR study.

Multivariable Mendelian Randomisation (MVMR) is a recently developed extension of MR that can be applied with either individual or summary level data to estimate the effect of multiple, potentially related, exposures on an outcome[3,4]. The three core assumptions that define a set of SNPs, G, as valid IV's for the purpose of an MVMR analysis are;

IV1: G must be strongly associated with *each* exposure given the other exposures included in the model;

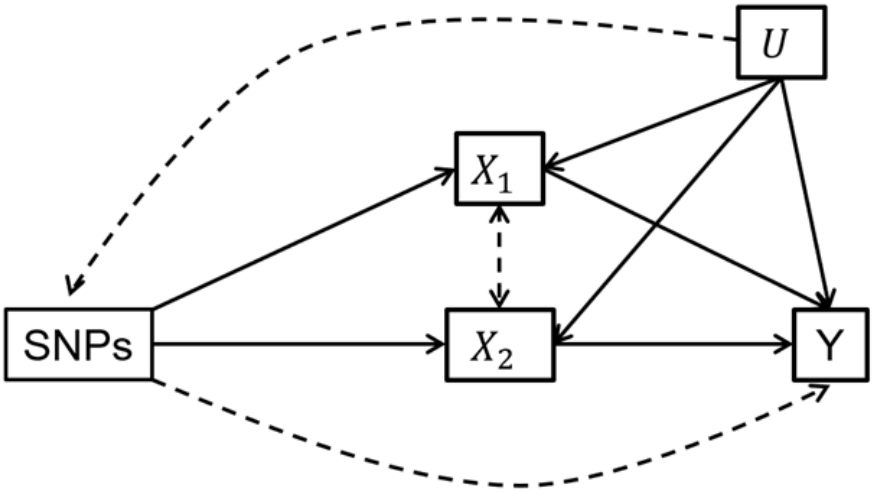
IV2: G is independent of all confounders of *any* of the exposures and the outcome; and

IV3: G is independent of the outcome given *all* of the exposures.[4].

These assumptions are illustrated in Fig 1. A violation of IV1 induces 'weak instrument bias' in the resulting estimates[5,6]. In a conventional (univariable) MR analysis, the definition of instrument strength is straightforward and unambiguous. Assumption IV1 can be tested with an F-statistic, which tests the association between the SNP and the exposure. When univariable MR analysis based on individual level data from a single sample, if the F-statistic is larger than the rule-of-thumb value of 10 then the SNPs are said to be a 'strong' instrument. We can then reject the null hypothesis that the instruments are weak in the sense that the bias of the MR estimate is equal to or greater than 10% of the observational (or ordinary least squares, OLS) association.[5,6]

In any MVMR analysis it is necessary that there are at least as many instruments as exposures and that this F-statistic is large for each exposure included, however this is no longer sufficient; the SNP's used as IV's also need to predict each exposure conditional on the other predicted exposures included in the estimation. This additional condition ensures that there is sufficient variation in association between the SNPs and each exposure, to avoid a problem of weak instrument bias in the MVMR model. Unlike in univariable MR, in MVMR weak instrument bias can bias the estimated effect of each exposure either towards or away from the null. This makes testing for weak instruments in any MVMR estimation particularly important.

Figure 1: Assumptions for a MVMR analysis



DAG illustrating the assumptions required for MVMR. Dashed lines represent associations that must not exist for the SNPs to be valid instruments for the set of exposures

With individual level data, weak instruments can be tested in MVMR using the Sanderson-Windmeijer conditional F-statistic, denoted F_{SW} [7,4]. Under weak instruments F_{SW} has the same distribution as the conventional F-statistic and so can be compared to the same critical values [5,6]. Therefore when testing for weak instruments, verifying that F_{SW} is greater than the rule-of-thumb of 10 means that we can reject the null hypothesis that the average bias of the MVMR estimates is at least 10% of the bias of the equivalent multivariable OLS estimates.

When individual level data on the genetic variants, exposure and outcome are not available two-sample MVMR can be conducted using summary data estimates of SNP-exposure and SNP-outcome associations. In two-sample MR, weak instruments bias the causal estimates towards

the null rather than the observational association[8]. In this paper we consider testing for weak instruments and estimation in the presence of weak instruments in the summary-data MVMR setting. Sanderson et al (2019) derived a Q statistic (Q_{x_j}) to test for underidentification (i.e. where the SNPs explain none of the variation in an exposure) in two-sample MVMR. We formally show in this paper that a transformation of this statistic has the same distribution as F_{SW} and therefore can also be compared to standard weak instrument critical values, or rule-of-thumb of F>10, to test for weak instruments in the two sample setting.

We then go on to consider horizontal pleiotropy in MVMR. Horizontal pleiotropy is a major threat to the validity of an MR analysis. It occurs when the SNPs have an effect on the outcome (either directly or through another exposure not included in the model) that is not via the exposure of interest, as illustrated by the dashed arrow from G to Y in Fig 1. This violates assumption IV3 and can lead to biased estimates of the causal effect of each exposure on the outcome from an MR analysis[9]. Horizontal pleiotropy can be either 'balanced', where the pleiotropic effects of the SNPs in the estimation are evenly distributed between having positive and negative effects on the outcome and so have no overall directional effect, or 'unbalanced' where on average these pleiotropic effects act in one direction on the outcome. IVW estimation and MVMR-IVW estimation are robust to balanced pleiotropy when the instruments are strong. However this no longer holds if the exposures are only weakly predicted by the SNPs. A number of methods currently exist for univariable MR estimation that are robust to pleiotropy under different assumptions.[10,11,12,13] MVMR can mitigate horizontal pleiotropy via known pleiotropic pathways through the inclusion of multiple exposures, however limited methods are available for pleiotropy robust MVMR models[4,14,15]. Furthermore, in the presence of weak instruments standard tests are increasingly likely to detect pleiotropy when in truth none is present. The major contribution of this paper is to extend weak instrument and pleiotropy robust estimation to two sample MVMR with an arbitrary number of exposures. Furthermore, we show that a heterogeneity statistic derived within this estimation procedure provides an exact test for the presence of pleiotropy in the presence of weak instruments. The methods presented here therefore provide the statistical framework for accurate and reliable MVMR model fitting, with potentially large numbers of exposures, in the presence of weak instruments and pleiotropy.

We apply our methods to determine whether particular subsets of metabolites can be strongly

predicted by 150 SNPs associated with at least one of 118 metabolites using data first presented by Kettunen et al 2016[16] and estimate the causal effect of those traits on Age related macular degeneration (AMD). The two-sample conditional F-statistic calculated for these data highlights that it is not possible to strongly predict multiple metabolites from the same subgroup despite each lipid fraction having a moderately high individual F-statistic and that any MVMR estimates including these is likley to be biased. Any analyst naively applying MVMR methods to such data without the correct diagnostic statistics to hand is in danger of generating poor quality results.

Finally we present an R package ('MVMR') that can conduct MVMR-IVW estimation and calculate all of the test statistics and estimators discussed in this paper.

2 A Test for Weak Instruments

Let $X = (X_1, X_2, ..., X_K)$ be a set of K exposure variables and let G be a set of L instruments $G = (G_1, G_2, ..., G_L)$. Define the $K \times L$ matrix of associations between each exposure and each instrument as;

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1L} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{K1} & \pi_{K2} & \dots & \pi_{KL} \end{pmatrix}, \tag{1}$$

where for example π_{32} represents the association between exposure 3 and SNP 2. Without loss of generality, testing whether the instrument set G can explain variation in a single exposure, X_1, conditional on all other exposures $(X_2, ..., X_K)$ is equivalent to testing whether model (2) below is identified

$$X_1 = \delta_{01} + \delta_1 X_{-1} + \epsilon_1 \tag{2}$$

$$X_m = \pi_{0m} + \sum_{j=1}^{L} \pi_{mj} G_j + \epsilon_m, \qquad m = 2, ..., K \tag{3}$$

Here: δ_{01} and each π_{0m} are scalar parameters; δ_1 is a $K-1$ vector of parameters, and ϵ_1 and ϵ_m are random error terms. Collecting $\pi_2, ..., \pi_K$ into a single $(K-1) \times L$ matrix, define Π_{-1} as the matrix Π minus its first row. This model considers only the exposures, and not the outcome, of the main estimation of interest as we wish to test whether the instruments explain any variation in X_1

over and above the variation explained in all of the other exposures. If this model is overidentified then the rank of Π_{-1} is strictly greater than $K-(L-1)$ and the instruments can strongly predict X_1 conditional on all other exposures included in the estimation.

In two sample summary data settings we do not directly observe exposures $X_1, ..., X_K$, only estimates for the $K \times L$ SNP-exposure associations that define $\hat{\Pi}$ the estimated value of Π obtained through regression of each exposure on each SNP. However we can use these association estimates to define an analogous formula to (2)

$$\hat{\pi}_1 = \delta_1 \hat{\Pi}_{-1} + v_1$$

The Q statistic for exposure 1 based on the summary data estimates can be written as;

$$Q_{x_1} = \sum_{j=1}^{L} \left(\frac{1}{\sigma^2_{x_1 j}} \right) \left(\hat{\pi}_{1j} - \tilde{\delta}_1 \hat{\Pi}_{-1j} \right)^2 \tag{4}$$

where the variance term $\sigma^2_{x_1 j}$ is given by;

$$\sigma^2_{x_1,j} = \tilde{\delta}^* \Sigma_{V,j} (\tilde{\delta}^*)'$$

Where j represents an individual SNP and π_{1j} and Π_{-1j} represent column j of π_1 and Π_{-1} respectively. $\tilde{\delta}^*$ is the K by 1 vector $(-1 \;\; \tilde{\delta}_2 \;\; \dots \;\; \hat{\delta}_K)$, and $\tilde{\delta}_k$ is a consistent estimator for δ_k, for example estimated through an inverse variance weighted least squares regression of $\hat{\pi}_1$ on $\hat{\Pi}_{-1}$. The matrix $\Sigma_{V,j}$ defines the covariance of the estimated effects of snp j on each of the exposures:

$$\Sigma_{V,j} = \begin{pmatrix} \sigma^2_{1,j} & \sigma_{12,j} & \cdots & \sigma_{1K,j} \\ \sigma_{12,j} & \sigma^2_{2,j} & \cdots & \sigma_{2K,j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K,j} & \sigma_{2K,j} & \cdots & \sigma^2_{K,j} \end{pmatrix} \tag{5}$$

If each $\hat{\pi}_{kj}$ is obtained separately via univariable regressions with an intercept, then the error terms

are obtained from the expressions:

$$\sigma^2_{k,j} = \frac{\left(G_j^T G_j\right)^{-1}}{n} \sum_{i=1}^{n} \hat{v}^2_{ki}, \quad \text{and} \quad \sigma_{km,j} = \frac{\left(G_j^T G_j\right)^{-1}}{n} \sum_{i=1}^{n} \hat{v}_{ki}\hat{v}_{mi}, \qquad k \neq m \tag{6}$$

Where $v_{k,i}$ and $v_{m,i}$ are the residual error terms for univariable regressions of SNP i on exposures k and m respectively. Under the null hypothesis that the instruments do not contain enough information to predict both exposure variables, Q_{x_1} will be asymptotically χ^2_{L-1} distributed where L is the number of SNPs in the estimation. Rejecting the null hypothesis indicates that the SNPs can predict X_1 conditional on X_2. Dividing the Q-statistic described above by the number of instruments, adjusted for the number of exposures, in the model gives a test statistic that is equivalent to the one sample conditional F statistic F_{SW}. Two-sample MVMR-IVW estimation is asymptotically equivalent to individual level two-stage least squares estimation and therefore this test statistic can be applied to test for weak instrument in two-sample MVMR in the same way as the conditional F-statistic for individual level data.[17]

$$\begin{aligned} F_{TS,k} =& \frac{Q_{x_k}}{L-(K-1)} \\ & \sim \frac{\chi^2_{(L-(K-1))}}{L-(K-1)} \end{aligned} \tag{7}$$

Where Q_{x_k} is the expression given in equation 4.

Critical values

Comparing this statistic to standard critical values from the F-distribution provides a test for a lack of identification. However, even if the genetic instruments explain some of the variation in the exposure they could still be 'weak'. In this case the estimates obtained from the MVMR estimation could still be considerably biased. The one sample conditional F-statistic (F_{SW}) has the same distribution as the Stock-Yogo weak instrument test[6]. Therefore we can apply its weak instrument critical values to identify weak instrument bias for univariable and multivariable two-sample MR[5,6,7]. The weak instrument critical values derived by Stock and Yogo (2005) for the bias of the 2SLS estimator relative to the OLS estimator are derived under the definition that the instruments are weak when the bias of the IV estimator relative to the OLS estimator is at least

10%. The measure of relative bias used is the squared bias of the IV estimator (β_{IV}) relative to the squared bias of the OLS estimator (β_{OLS}). This is given by the equation;

$$B^2 = \frac{(E\hat{\beta}_{IV} - \beta)'\Sigma_X(E\hat{\beta}_{IV} - \beta)}{(E\hat{\beta}_{OLS} - \beta)'\Sigma_X(E\hat{\beta}_{OLS} - \beta)}$$

Where $\Sigma_X = plim\frac{1}{n}X'X$ and X here represents the $n \times K$ matrix of all of the exposures included in the estimation, n is the sample size. Calculating the bias in this way standardises the exposures X so they are orthogonal and have unit standard deviation. However it means that the bias of the estimated effect of any particular exposure may differ from 10% and the bias of any particular exposure could act in the opposite direction to the bias of the model as a whole. If F_{TS} is larger than the relevant Stock-Yogo critical value we can reject the null hypothesis that the exposure is only weakly predicted by the instruments. These critical values have only been derived for models including up to 30 instruments, therefore in Table 1 we provide critical values for a larger range of instruments to test for a 5%, 10% or 20% relative bias. These critical values are often approximated to a rule of thumb of $F > 10$ to test a null hypothesis that the bias is at least 10% of the bias of the OLS estimator. The critical values given above also show that the rule of thumb of 10 is slightly smaller than the true critical value for this test and would lead to the null hypothesis being rejected more frequently. The two sample F_{TS} statistic tests the bias of the model as a whole, this means that the sign of the bias of an individual causal parameter may differ from that of the model's bias, which is averaged across all of its constituent parameters. It also indicates that some weakly predicted exposures could be biased away from the null hypothesis.

Table 1: Critical values for conditional weak instrument tests.

	Relative bias		
k_Z	5%	10%	20%
25	21.37	11.44	6.19
50	21.26	11.14	5.86
100	21.02	10.84	5.64
200	20.79	10.61	5.46
300	20.62	10.52	5.38
400	20.56	10.45	5.32
500	20.50	10.40	5.29

3 Weak Instrument Robust Two-sample MVMR

Estimation in the presence of weak instruments

In the presence of weak instruments, standard inverse variance weighted estimation of the MVMR mode, which we refer to as MVMR-IVW, is biased. The LIML estimator has previously been proposed as an alternative estimator for individual-level MR as it is less biased when there are many weak instruments[18]. In the two-sample summary data setting, Bowden and colleagues[19] and Zhao and colleagues[20] show that weak instruments can be effectively mitigated through minimisation of an appropriate heterogeneity statistic using weights that account for the variance of the SNP-exposure associations is analagous to one-sample LIML estimation. It gives results that are substantially less biased than conventional regression based IVW estimates in the presence of a non-zero causal effect. The weak instrument robust estimation proposed by Bowden and colleagues can be extended to the MVMR setting as a minimisation of;

$$Q_A = \sum_{j=1}^{L} \left(\frac{1}{\sigma^2_{A,j}} \right) \left(\hat{\Gamma}_j - \beta' \hat{\pi}_j \right)^2 \tag{8}$$

over β. Where β is a vector of causal parameters (to be estimated), $\hat{\Gamma}_j$ is the estimated effect of SNP j on the outcome, $\hat{\pi}_j$ is a vector of effects of SNP j on each exposure included in the estimation (i.e. a column of the matrix Π) and;

$$\sigma^2_{A,j} = \sigma^2_{y,j} + \beta' \Sigma_{V,j} \beta. \tag{9}$$

Here, $\sigma^2_{y,j}$ is the variance of the estimated effect of the SNPs on the outcome, and $\Sigma_{V,j}$ is the variance-covariance matrix defined in equation (5). This is equivalent to minimisation of the Q_A statistic to test for heterogeneity described in Sanderson et al 2019[4] extended to a model with more than two exposures. We label estimates for β obtained in this manner as $\hat{\beta}_Q$. The standard MVMR-IVW estimate is vunerable to weak instrument bias because instead of minimising Q_A in (8) using the full weights defined in (9) it incorrectly assumes that $\sigma^2_{A_j} = \sigma^2_{y_j}$. This ignores the component of variation from $\beta' \Sigma_j \beta$ and is only valid if either all elements of β are zero or Σ_j is negligable in comparison to σ^2_{yj}.

Testing for pleiotropy in the presence of weak instruments

Horizontal pleiotropy - where genetic variants influence the outcome through multiple phenotypes can lead to a violation of the IV assumptions if they are not included as exposures in the MVMR estimation. Under the assumption that not all the SNPs included in the estimation have a pleiotropic effect on the outcome through the same pathway, this will lead to greater variation in the estimated causal effect of the exposures on the outcome than would be expected by chance. This excess heterogeneity can be reliably tested for using the minimised Q_A statistic. More formally if all SNPs used in the MVMR analysis are valid instruments, in the sense that they identify a common set of causal parameters β, we would expect the Q_A statistic in (6) evaluated at $\beta = \hat{\beta}_Q$ to follow a Chi-squared distribution with L-K degrees of freedom. Crucially, the test is exact in the sense that it will achieve its nominal type I error rate, even in the presence of weak instruments[21]. The standard Q-statistic used to generate the MVMR-IVW estimate by setting $\sigma^2_{A,j} = \sigma^2_{y,j}$, referred to here as Q_{IVW}, will generally have an inflated type 1 error rate (i.e. will detect pleiotropy too often when none is present) unless all $\beta'\Sigma_j\beta$ terms are negligible.

Estimation with pleiotropic and weak instruments

Estimation of β through minimisation of (8) will give estimates of the direct effect of each exposure on the outcome that are robust to weak instruments. However, these estimates will still be biased in the presence of pleiotropy. In order to account for heterogeneity due to pleiotropy, we extend the estimation of β by adding a pleiotropy variance parameter τ^2 to the multivariable Q estimation and finding the joint value of (β, τ^2) which minimises;

$$\sum_{j=1}^{L} \left(\frac{1}{\sigma^2_{A,j}} \right) \left(\hat{\Gamma}_j - \beta' \hat{\pi}_j \right)^2 - (L - K) = 0$$

$$\sigma^2_{A,j} = \sigma^2_{y,j} + \beta' \Sigma_j \beta + \tau^2$$

subject to;

$$\frac{\partial \sum_{j=1}^{L} \left(\frac{1}{\sigma^2_{A,j}} \right) \left(\hat{\Gamma}_j - \beta' \hat{\pi}_j \right)^2}{\partial \beta} = 0$$

We refer to the causal estimates derived in this way as $\hat{\beta}_{Q,het}$. This is a extension of the method described in Bowden et al 2018 for univariable MR to the MVMR setting.[19] This method will

account for balanced pleiotropy which biases the MVMR-IVW estimates further in the presence of weak instruments by accounting for excess heterogeneity in the per SNP estimated effects that is not related to the variance in the SNP-exposure associations or SNP-outcome associations. It will not however account for directional pleiotropy where the pleiotropic effects of the SNPs on the outcome all, or mostly, act in one direction to either increase or decrease the outcome. However, it is possible to look at the individual contribution of each SNP to Q_A to identify the largest outliers. If a small number of SNPs are observed to have a large effect on Q_A they can potentially be removed as a sensitivity analysis and the MVMR model re-estimated without them.

Confidence intervals for estimated effects

Estimation of β and τ^2 through minimisation of Q_A, does not provide readily available and reliable standard errors. We therefore suggest that standard errors are obtained, and confidence intervals calculated, through a Jackknife procedure.

We propose the use of Jackknife rather than a bootstrap as with a moderate number of SNPs the repeated sampling in a bootstrap can lead to very weak instruments in any particular iteration even when the model has relatively strong instruments as a whole. A jackknife procedure estimates the model leaving out each SNP in turn and then calculates the standard deviation of the effect estimate from these results. As each iteration includes all but one of the SNPs and includes each SNP only once this is unlikely to be affected by weak instruments due to the exclusion of some SNPs. When the number of SNPs used in the estimation is very small neither a Jackknife or bootstrap approach will calculate appropriate standard errors however many applications of MVMR include 100 - 200 SNPs as instruments and with this number of SNPs a jackknife approach will be feasible.

4 Estimation of Σ_{Vj}

So far we have assumed that the pairwise covariance between a set of SNP's estimated association with any two exposures is known for all exposures and all SNPs. However, this data is not generally reported by GWAS summary statistics. Similarly it would not be feasible for these studies to report this data due to the large number of potential covariances that could be required for all potential future MVMR analyses. Excluding these covariances will give the correct estimation only under the global null ($\beta = 0$).

Therefore, in this section we suggest three different solutions for dealing with the lack of covariances in the GWAS summary results in order to estimate $\sigma_{km,j}$: the covariance between $\hat{\pi}_{k,j}$ and $\hat{\pi}_{m,j}$ with respect to exposure, k, exposure, m ($k \neq m$) and SNP j which form the elements of $\Sigma_{V,j}$.

Estimate $\sigma_{km,j}$ from the individual level data If some or all of the individual level data that was used in the GWAS to estimate the SNP - exposure associations is available then the covariances for the effect of each SNP on each exposure can be calculated from equation 6.

Estimate the phenotypic correlation between the exposures from individual level data The covariance for each SNP can then be approximated as;

$$\sigma_{km,j} = \rho_{km}\sigma_{k,j}\sigma_{m,j} \tag{10}$$

where ρ_{km} is the correlation between X_k and X_m (or phenotypic correlation). σ_{kj} and σ_{mj} are the standard error for the effect of SNP j on exposures k and m respectively. Although ideally this information would be calculated from the data used for the GWAS study, ρ_{km} could also be estimated from only part of the data used in the GWAS or from an alternative dataset which is thought to have a similar structure.

Estimate the effect of the SNPs on each exposure from separate samples Estimating the effect of the SNPs on each exposure in this manner means that the covariances will be zero and so excluding this information will not affect the statistics calculated. For an MVMR analysis involving K exposures, this would require $K + 1$ separate samples and so is likely to only be practicable in a limited number of cases.

In any given scenario some of these solutions may be impossible (due to a lack of data) and of the solutions that are possible, one may be the most reasonable. We suggest that estimation of ρ_{km} from phenotypic data, from which the appropriate covariances can then be calculated, is likely to be the most feasible and appropriate approach in many cases.Under the assumption that each SNP explains a small proportion of the variation in the exposure, the accuracy of the estimate of $\sigma_{km,j}$ will depend on the accuracy of the estimate of ρ_{km}. Therefore when ρ_{km} is estimated from data that does not closely match that used to estimate the SNP exposure associations exploration

of how sensitive F_{TS} and $\beta_{Q,het}$ are to that estimate should also be conducted. This could be done through estimation of F_{TS} and $\beta_{Q,het}$ at the limits of or across the range of reasonable values of ρ_{km}. These results should then be used to determine whether the interpretation of the results changes over plausible values of ρ_{km}.

5 Simulation Results

To illustrate the methods presented so far give here results from simulating and fitting MVMR models with 200 SNPs and either 2 or 3 exposures.

MVMR model with two exposures

Firstly, we simulated a MVMR model with 2 exposures and 200 SNPs. The SNP-exposure associations where constructed in two ways; firstly so that each exposure was individually and conditionally weakly predicted by the set of SNPs (i.e. weak instruments) and secondly so that the exposures were strongly individually predicted, but weakly conditionally predicted by the set of SNPs (i.e. conditionally weak instruments). In each case the association of each SNP with the exposure was drawn from a uniform distribution with the range of association selected to maintain the desired overall instrument strength. All of the SNPs were associated with both exposures, for the weak instruments there was no correlation between the association between each SNP and each exposure. Conditionally weak instruments were generated by increasing the total strength of the instruments but introducing correlation between the effect of each SNP on each of the exposures following the structure of weak instrument asymptotics first introduced by Staiger and Stock (1997) [5]. This reflects a scenario where examination of standard F-statistics for each exposure would not identify weak instruments. The exposures were simulated to both have a direct effect on the outcome and balanced pleiotropy was introduced to the model through a direct effect of the SNPs on the outcome. Pleiotropic effects were generated from a normal distribution with zero mean. A confounder of both exposures and the outcome was also included. The covariance parameter $\sigma_{i,j},\ i \neq j$ was estimated from calculation of the phenotypic correlation between X_1 and X_2 as described in section 4. The set up of this model is illustrated in Fig. 2 and results from the simulation are given in Table 2. Results for the same model without the pleiotropic effect of the SNPs on the outcome are given in Supplementary Table S.1.

Results from this simulation show that the two-sample conditional F statistic F_{TS} reliably estimates the strength of the instruments and is equivalent to the conditional F statistic calculated from the individual level data F_{SW} when the correlation between the exposures is used to estimate the covariance between the effect of each SNP on each exposure. These results also show that although $\hat{\beta}_Q$ does not reliably estimate the effect of the exposure on the outcome in the presence of balanced of pleiotropy, $\hat{\beta}_{Q,het}$ which allows for this additional heterogeneity does. This decrease in bias in $\hat{\beta}_{Q,het}$ compared to $\hat{\beta}_{MVMR-IVW}$ when the instruments are weak comes at the cost of increased standard errors, reflecting the (true) lower level of information in the model. Supplementary Table S.1 shows that allowing for heterogeneity when it is not present does not increase the standard error of the $\hat{\beta}_{Q,het}$ estimates relative to the standard error of the $\hat{\beta}_Q$ estimate. Table 2 also gives F_{TS} and $\beta_{Q,het}$ estimated without accounting for σ_{km}, labelled $F_{TS,0}$ and $\hat{\beta}_{Q,het,0}$ respectively. This imposes the assumption that $\sigma_{km} = 0, \quad k \neq m$, but not the assumption that $\sigma_k^2 = 0$ and so is a point between standard MVMR-IVW estimation and $\hat{\beta}_{Q,het}$. These results also show that in the presence of conditionally weak instruments, when there is correlation between the effect of the SNPs on each exposure, if these correlations are not taken into account $F_{TS,0}$ does not reliably test the strength of the instruments and $\hat{\beta}_{Q,het,0}$ produces biased estimates of the effect of each exposure on the outcome.

Three exposure model

Next we simulated summary data for three exposures and 200 SNPs. Each of the exposures was simulated to have a direct effect on the outcome. All of the SNPs included in the estimation are associated with every exposure. The effect of the SNPs on the second exposure was uncorrelated with the effects on the first or third exposures. However, the effect of the SNPs on the first and third exposures were correlated, so that the third exposure was only weakly predicted by the SNPs conditional on the first exposure (and therefore the first exposure is weakly predicted conditional on the third exposure). This set up means that when only the first two exposures are included in the estimation there is directional pleiotropy present, however when all three exposures are included there is potential weak instrument bias. When the two exposures are included they each have mean conditional F-statistics of 45 whereas in the model with three exposures included exposure 1 has a mean conditional F-statistic of 6.5 and exposure 3 has a mean conditional F-statistic 3.2.

Figure 2: Model simulated in Table 2

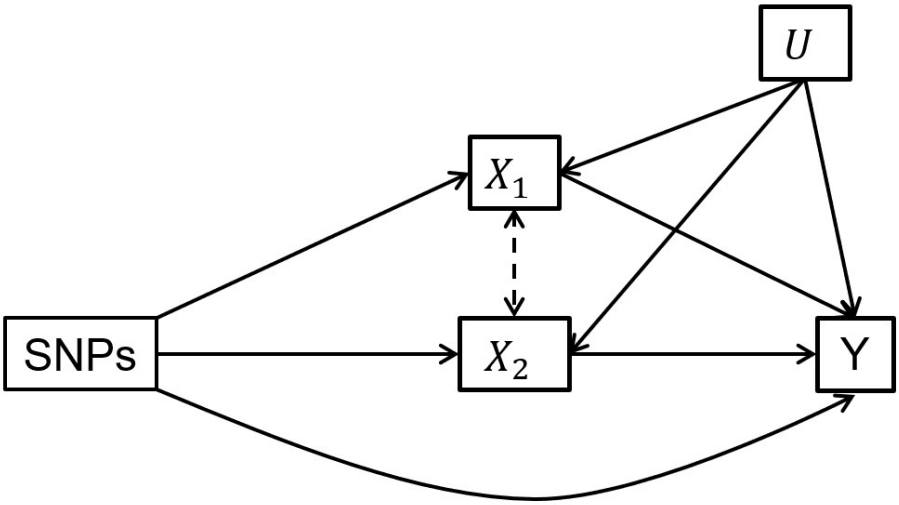


Table 2: Simulation results for models with heterogeneity: two exposures, 200 SNPs

	Weak instruments		Conditionally weak instruments	
	x_1	x_2	x_1	x_2
One-sample estimation with individual level data				
$\hat{\beta}_{OLS}$	1.09	-0.049	0.78	-0.48
	(0.033)	(0.033)	(0.029)	(0.026)
$\hat{\beta}_{IV}$	0.585	-0.283	0.548	-0.333
	(0.533)	(0.533)	(0.311)	(0.226)
F	8.80	8.80	1602.81	3107.5
	(0.61)	(0.62)	(107.67)	(208.06)
F_{SW}	3.40	3.40	9.75	9.78
	(0.360)	(0.360)	(0.94)	(0.95)
Two-sample estimation with covariances				
$\hat{\beta}_{IVW}$	0.352	-0.128	0.469	-0.276
	(0.541)	(0.541)	(0.316)	(0.228)
$\hat{\beta}_{Q}$	$-7.7x10^3$	$6.7x10^3$	$-6.6x10^5$	$4.7x10^5$
	$(1.2x10^5)$	$(1.0x10^5)$	$(2.3x10^6)$	$(1.6x10^6)$
$\hat{\beta}_{\mathbf{Q,het}}$	**0.487**	**-0.246**	**0.519**	**-0.313**
	(0.777)	**(0.778)**	**(0.350)**	**(0.253)**
$\mathbf{F_{TS}}$	**3.35**	**3.35**	**9.13**	**9.15**
	(0.348)	**(0.347)**	**(0.814)**	**(0.819)**
Two-sample estimation without covariances				
$\hat{\beta}_{IVW}$	0.352	-0.128	0.469	-0.276
	(0.541)	(0.541)	(0.316)	(0.228)
$\hat{\beta}_{Q}$	$-6.8x10^3$	$6.0x10^3$	$-6.0x10^5$	$4.3x10^5$
	$(1.1x10^5)$	$(9.5x10^4)$	$(6.5x10^5)$	$(4.8x10^5)$
$\hat{\beta}_{Q,het}$	0.499	-0.260	$-4.5x10^5$	$3.2x10^5$
	(0.802)	(0.803)	$(1.5x10^6)$	$(1.1x10^6)$
F_{TS}	3.17	3.17	0.45	0.45
	(0.337)	(0.336)	(0.054)	(0.054)

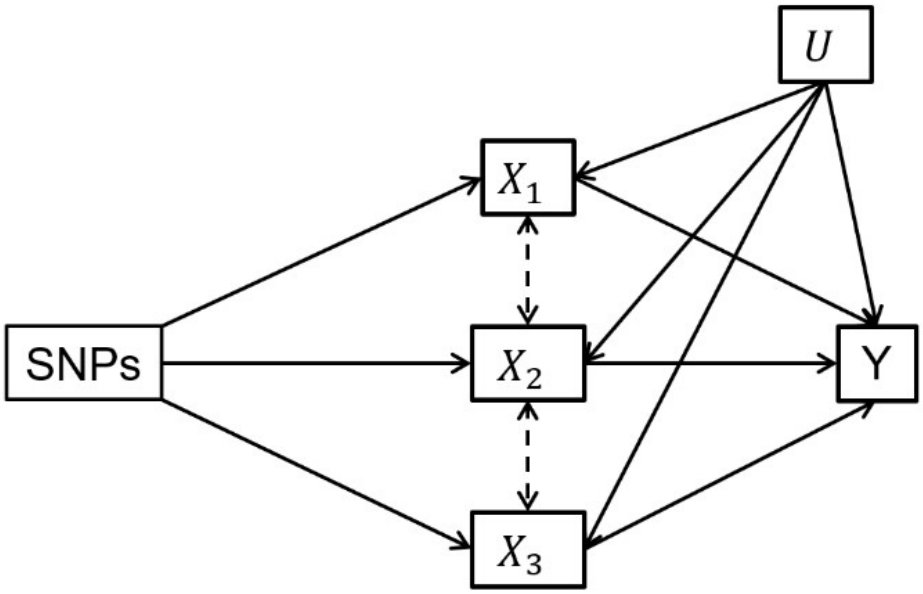
$\beta_1 = 0.5$, $\beta_2 = -0.3$
4,000 repetitions, 20,000 observations per repetition
Covariances estimated from the phenotypic correlation between each exposure.
Weak instruments shows a scenario where the exposures are individually weakly predicted by the SNPs. Conditionally weak instruments gives a scenario where the exposures are strongly predicted by the SNPs individually but are each weakly predicted by the SNPs conditional on the other exposure.

When three exposures are included in the model exposure 2 is still strongly predicted with a mean conditional F-statistic of 17.9. The model under which the data was generated is illustrated in Fig.

3 and results are given in Table 3.

We give results from estimation of the model firstly including only two exposures, x_1 and x_2, and including all three exposures. These results show that when only two exposures are included in the model all methods of estimating β_1 and β_2 are biased by the directional pleiotropy present in the model. When all three exposures are included in the model the MVMR-IVW estimates are biased due to the presence of weak instruments. However, estimation of $\hat{\beta}_Q$ through minimisation of Q_A gives unbiased estimates of the effect of each exposure.

Figure 3: Model simulated in Table 3



Heterogeneity Testing

Table 4 gives the rejection rates when using Q_{IVW} and Q_A to test for pleiotropy for the model considered in Figure 2. In addition, we show rejections rates using a third heterogeneity statistic that attempts to improve $Q_{\sigma_y^2}$ by extending the weights so they take the form $\sigma_y^2 + \beta' \Sigma_j \beta$. We call this heterogeneity statistic $Q_{IVW,up}$. These extended weights are calculated using a multivariable

Table 3: Simulation results for a model with three exposures.

	Two exposures included in estimation		Three exposures included in estimation		
	x_1	x_2	x_1	x_2	x_3
One-sample estimation with individual level data					
$\hat{\beta}_{OLS}$	0.837	-0.065	0.667	-0.176	1.418
	(0.020)	(0.019)	(0.020)	(0.018)	(0.012)
$\hat{\beta}_{IV}$	0.626	-0.244	0.466	-0.314	0.912
	(0.018)	(0.018)	(0.017)	(0.014)	(0.064)
F	236.2	235.13	236.22	235.13	14.50
	(15.43)	(13.67)	(15.43)	(13.67)	(0.89)
F_{SW}	78.49	78.37	6.62	19.81	3.17
	(7.10)	(6.98)	(1.13)	(5.38)	(0.29)
Two-sample estimation with covariances					
$\hat{\beta}_{IVW}$	0.611	-0.228	0.523	-0.277	0.502
	(0.021)	(0.021)	(0.027)	(0.021)	(0.101)
$\hat{\beta}_{Q}$	0.626	-0.246	0.500	-0.301	0.703
	(0.022)	(0.021)	(0.034)	(0.024)	(0.149)
$\hat{\beta}_{\mathbf{Q,het}}$	**0.624**	**-0.246**	**0.499**	**-0.301**	**0.705**
	(0.022)	**(0.021)**	**(0.035)**	**(0.024)**	**(0.154)**
$\mathbf{F_{TS}}$	**45.01**	**44.97**	**6.58**	**17.94**	**3.23**
	(2.39)	**(2.35)**	**(1.09)**	**(4.32)**	**(0.29)**

$\beta_1 = 0.5$, $\beta_2 = -0.3$, $\beta_3 = 0.7$
4,000 repetitions, 20,000 observations per repetition
Covariances estimated from the phenotypic correlation between each exposure. Two exposures included in estimation refers to estimation of the model including only exposures 1 and 2. Three exposures included in estimation includes exposures 1, 2 and 3.

extension of the iterative estimation described in Bowden et al (2019).[19] These results show that when there is no heterogeneity the null hypothesis is over rejected by both Q_{IVW} and $Q_{IVW,up}$. Although the iterative updating improves on standard estimation it does not fully correct for the over rejection due to weak instruments.[19] Estimation of Q_A using direct minimisation controls the type 1 error and when the null hypothesis is true, i.e. when there is no heterogeneity this test statistic rejects approximately 5% of the time.

6 Application

In this section we illustrate the use of the methods described above through an application to the estimation of the effect of multiple metabolites to age-related macular degeneration (AMD).

Table 4: Estimation of Q_A.

	Weak instruments				Conditionally weak instruments			
	$\tau^2 = 0$		$\tau^2 = 0.5$		$\tau^2 = 0$		$\tau^2 = 0.5$	
	Estimate	Rej. Rate	Estimate	Rej. Rate	Estimate	Rej. Rate	Estimate	Rej. Rate
Q_{IVW}	228.37	41.9%	13366.76	100%	249.71	75.4%	13032.82	100%
	(23.04)		(678.27)		(24.85)		(770.59)	
$Q_{IVW,up}$	206.20	12.9%	11645.15	100%	201.36	7.6%	11593.22	100%
	(21.39)		(1699.61)		(20.53)		(1338.27)	
Q_A	197.16	4.5%	576.93	100%	197.74	5.2%	1788.42	100%
	(19.83)		(53.92)		(19.85)		(224.66)	

4,000 repetitions, 20,000 observations per repetition
Covariances estimated from the phenotypic correlation between each exposure.

AMD is disease that causes loss of central vision and is a leading cause of blindness[22]. Elevated lipid serum levels have previously been associated with increased risk of AMD[23]. We use data from a Genome-wide association study (GWAS) of 118 metabolites by Kettunen et al 2016[16] as our exposure and from a GWAS of AMD as our outcome[24]. The GWAS data for our exposures included 150 SNPs that were genome-wide signification for at least one of the metabolites. Previous studies have implicated HDL as being causal for AMD[25,26,27]. In this analysis we illustrate the issues with weak instrument bias that can arise from including multiple highly related traits in one MVMR estimation.

The GWAS data included 118 potential metabolite exposures. For the purposes of illustration we restricted the analysis to 14 metabolites moderately well predicted by a large number of SNPs. Specifically we selected the 13 metabolites that had 42 or more SNPs with an F-statistic greater than 5 associated with them in our data. From the 150 SNPs included in the data we retained all SNPs which were associated with at least one of our selected exposures with an F statistic greater than 5. This gave us 78 SNPs associated with our 13 metabolites for our analysis. From this data we considered 6 different models to estimate and for each one obtained the MVMR-IVW effect estimates and investigated whether the SNPs included as instruments could conditionally predict the exposures in that model. The models considered were (a) all selected metabolites (b) - (e) subgroups of metabolites grouped by lipid fraction type and (f) a subgroup including one metabolite from each group included in (b) - (e).

Table 5 gives results for the estimation of model (a) including all of the selected metabolites.

This table also reports; the mean individual F-statistic for the SNPs associated with each metabolite (F_{ind}), the mean F-statistic across all of the SNPs included in the analysis for each metabolite (F_{all}) and the conditional F-statistic for each metabolite (F_{TS}). The correlation between the metabolites, required to calculate F_{TS}, was not available from the GWAS data used here. We therefore calculated these using external data on the same metabolites from the Avon Longitudinal Study of Mothers and Children (ALSPAC)[28,29]. A description of the ALSPAC study is given in the supplementary material. The F-statistics and conditional F-statistics presented for the model including all metabolites show that although each metabolite is strongly predicted by the SNPs associated with it the conditional F-statistics for each exposure are very small and therefore the effect estimates are subject to weak instrument bias.

Table 5: MVMR estimates of a range of metabolites on AMD, all metabolites included in one MVMR estimation

		Estimate	Std. Error	P-value	F	F_{TS}
ApoB	ApoB	1.673	0.693	0.019	10.82	0.197
IDL	IDL.PL	-4.456	0.969	<0.001	11.84	0.011
	IDL.P	6.481	3.396	0.061	11.76	0.626
	IDL.TG	0.437	1.391	0.754	11.04	0.003
LDL	L.LDL.L	-8.695	8.376	0.303	11.15	0.001
	L.LDL.P	5.223	11.125	0.640	11.34	0.001
	M.LDL.P	1.794	2.360	0.450	10.56	0.011
Small VLDL	S.VLDL.PL	1.054	1.530	0.493	8.62	0.029
	S.VLDL.C	1.346	1.617	0.408	8.88	0.005
	S.VLDL.FC	-1.270	1.331	0.343	8.75	0.019
Very Small VLDL	XS.VLDL.L	-6.655	1.982	0.001	10.67	0.027
	XS.VLDL.P	4.866	1.668	0.005	10.19	0.048
	XS.VLDL.TG	-2.384	1.819	0.195	9.14	0.022

F is the mean F-statistic across all SNPs included in the estimation and is the univariable F-statistic for instrument strength. F_{TS} is the conditional F-statistic accounting for the association between each SNP and all of the other exposures included in the estimation.
78 SNPs included in the estimation. ApoB is associated with 48 SNPs, IDL.PL, L.LDL.L, L.LDL.p, M.LDL.P, S.VLDL.PL and XS.VLDL.L are each associated with 43 SNPs, IDL.P, IDL.TG, S.VLDL.C, S.VLDL.FC, XS.VLDL.P adn XS.VLDL.TG are each associated with 42 SNPs.

Table 6 gives the same results for the estimation for each sub group of metabolites (IDL, LDL, Small VLDL and Very Small VLDL). These results show that, with the exception of IDL.PL and S.VLDL.PL, none of the metabolites are strongly conditionally predicted by the SNPs within their subgroup. For our last analysis we included one metabolite from each group as exposures in our MVMR estimation. Table 7 gives results for this set of exposures. Although the exposures

here are jointly moderately strongly predicted by the set of SNPs the conditional F-statistics for each exposures are still between 4.2 and 8.3 indicating that there is likely to be some weak instrument bias. In Table 8 we re-estimate this final MVMR model using our weak instrument robust estimators presented earlier. The results from this approach suggest that in our final model the initial MVMR-IVW estimates may be biased towards the null due to weak instruments. Q_A for this model is 118, the critical value at a 5% level of significance for a chi-squared distribution with 64 degrees of freedom is 84.7. It therefore indicates potential pleiotropy and we consider the $\hat{\beta}_{Q,het}$ to be the most appropriate estimates in this case. Comparison of these results to those obtained from model (a) including all of the metabolites shows the potential for weak instruments to bias results of a summary-data MVMR away from the null as well as towards the null. For 3 of the 4 metabolites included in both models the effect estimates in the final model are much closer to zero than the results in the model including all of the metabolites. The results from this analysis suggest that none of the metabolites considered are causally associated with AMD but that the standard MVMR-IVW estimates for the final model were biased due to both weak instruments and pleiotropic effects of the SNPs on the outcome. This null result is consistent with other results using an alternative method to analyse the same data which found that HDL (not included in this analysis) was the only metabolite that was causally associated with AMD[27].

7 Software

We have written an R package MVMR which facilitates the implementation of MVMR estimation and corresponding sensitivity analyses. The package requires summary data on instrument-exposure and instrument-outcome associations, as well as information on the pairwise covariances of the error in the estimated association between each SNP and each pair of exposures. As these covariances are often not available the software can be implemented in three ways; estimating the covariances from individual level data, approximating the covariances from the phenotypic correlation between the exposures or assuming that these covariances are zero.

Workflow

Fitting and interpreting MVMR using the methods described in this paper, including tests for instrument strength and horizontal pleiotropy, is performed using a five-step procedure. Initially,

Table 6: MVMR estimates of a range of metabolites on AMD, estimated by subgroup

	Estimate	Std. Error	P-value	F	F_{TS}
ApoB	ApoB				
IDL; 54 SNPs					
IDL.PL	-1.338	1.091	0.226	16.05	1.23
IDL.P	1.864	1.231	0.134	16.12	1.24
IDL.TG	-0.926	0.398	0.024	14.97	2.23
LDL; 46 SNPs					
L.LDL.L	3.707	4.341	0.398	17.58	0.019
L.LDL.P	-4.781	3.484	0.177	73.83	0.023
M.LDL.P	0.896	1.443	0.538	16.55	0.063
Small VLDL; 50 SNPs					
S.VLDL.PL	-0.513	1.021	0.617	12.38	11.65
S.VLDL.C	-0.372	0.858	0.667	12.42	4.75
S.VLDL.FC	0.506	1.298	0.698	12.51	5.39
Very Small VLDL; 53 SNPs					
XS.VLDL.L	-1.651	1.863	0.380	14.64	0.174
XS.VLDL.P	-0.105	0.533	0.845	12.50	0.916
XS.VLDL.TG	1.395	2.112	0.512	13.99	0.176

F is the mean F-statistic across all SNPs included in the estimation and is the univariable F-statistic for instrument strength. F_{TS} is the conditional F-statistic accounting for the association between each SNP and all of the other exposures included in the estimation.

Table 7: MVMR-IVW estimates of a range of metabolites on AMD including one exposure from each subgroup

	Estimate	Std. Error	P-value	F	F_{TS}
XS.VLDL.P	-0.778	0.958	0.420	11.26	4.23
S.VLDL.PL	0.051	0.347	0.385	9.48	5.68
L.LDL.L	0.356	0.231	0.154	12.19	8.22
IDL.TG	0.067	0.761	0.969	12.21	6.15

69 SNPs

F is the mean F-statistic across all SNPs included in the estimation and is the univariable F-statistic for instrument strength. F_{TS} is the conditional F-statistic accounting for the association between each SNP and all of the other exposures included in the estimation.

summary data should be provided, including a covariance matrix for the effect of the genetic variants on each exposure. As such covariances are not conventionally reported in publicly available data, two functions `snpcov_mvmr()` and `phenocov_mvmr()` can be used to generate the covariance matrix. The function `snpcov_mvmr()` estimates the covariance terms directly from individual level data, whilst `phenocov_mvmr()` uses the phenotypic correlation and summary data (input by the user) to generate estimates of the covariances.

Table 8: Weak instrument robust estimates of a range of metabolites on AMD including one exposure from each subgroup

	$\hat{\beta}_Q$			$\hat{\beta}_{Q,het}$		
	Est.	Std. Error	p-value	Est.	Std. Error	p-value
XS.VLDL.P	-5.008	3.774	0.185	-2.071	1.447	0.152
S.VLDL.PL	0.957	0.940	0.309	0.300	0.528	0.570
L.LDL.L	1.534	0.645	0.017	0.728	0.613	0.235
IDL.TG	2.490	2.614	0.341	0.803	1.437	0.576

69 SNPs
$\hat{\beta}_Q$ gives the estimate obtained by minimisation of Q, $\hat{\beta}_{Q,het}$ gives the estimate obtained by minimisation of Q allowing for balanced pleiotropy.

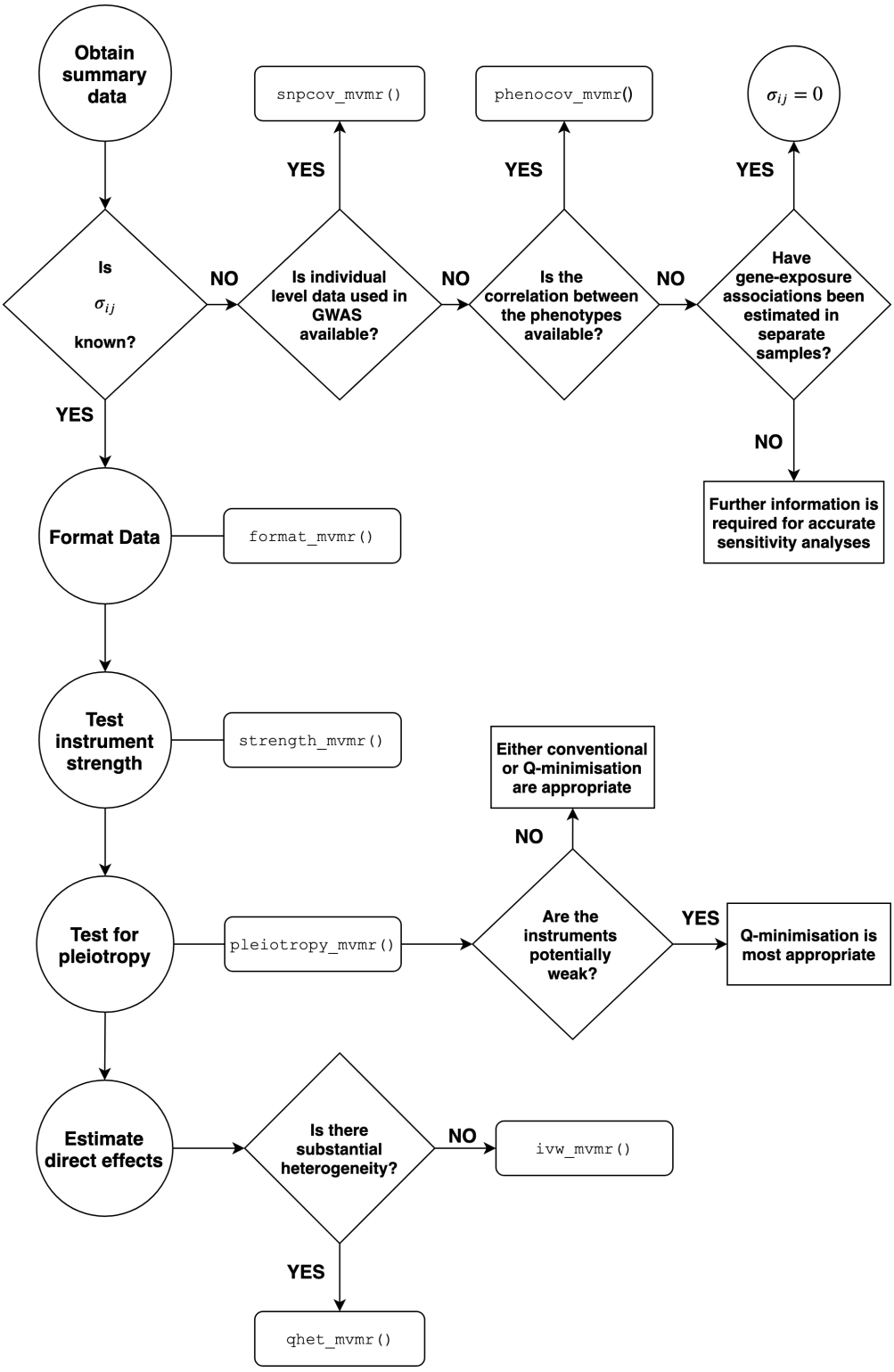
As a second stage, the summary data is reformatted using the function `format_mvmr()` into a data frame which is subsequently used as the input for estimation and sensitivity analyses. We then provide the functions `strength_mvmr()` to evaluate instrument strength using the two sample conditional F-statistic described in Section 2. Tests for horizontal pleiotropy are performed using `pleiotropy_mvmr()`, performing both standard and Q-minimisation approaches simultaneously (see section 3 for more details). Finally, causal effects can be estimated using two different approaches; fitting an inverse variance weighted (IVW) MVMR model using `ivw_mvmr()`and minimising the Q-statistic allowing for heterogeneity using `qhet_mvmr()`. Each step in the MVMR workflow is illustrated in Figure 4. The MVMR package is available to download at https://github.com/WSpiller/MVMR/. The package also includes a detailed tutorial demonstrating functionality of the package in an analyses of the effects of lipid fractions upon systolic blood pressure using data from the Global Lipids Genetics Consortium and UK Biobank.

8 Discussion

In this paper we develop a general statistical framework for conducting two sample MVMR analyses for an arbitrary number of exposures in the presence of weak instrument bias and pleiotropy. The methods presented here give ways to test for weak instruments in two-sample MVMR and to robustly test for heterogeneity due to pleiotropy in the presence of moderately weak instruments. We additionally give a method to estimate causal effects in the presence of moderately weak instruments which is robust to balanced pleiotropy.

Weak instruments are a potential issue in many applications where estimating direct effects of

Figure 4: Workflow for MVMR R package



multiple exposures using MVMR is preferred over univariable MR analyses , which are thought to be likely to be affected by directional pleiotropy[30,31,32,33,34]. MVMR approaches are also used to gague the extent to which one exposure mediates the effect of another on the outcome[35,36].Any application of MVMR will be biased by conditionally weak instruments and, as illustrated by our application, this can occur even when the genetic variants strongly predict each exposure individually. Therefore, the methods presented here are important as they provide a way to identify and correct for weak instruments in two-sample MVMR estimation.

The F_{TS} statistic described here is calculated using estimates of $\hat{\delta}$ calculated from an IVW estimation of the effect of $\hat{\pi}_{-k}$ on $\hat{\pi}_k$. An alternative method of estimation, equivalent to that described for estimation of β, is to directly minimise its constituent Q_{x_k} to obtain LIML estimates for δ[19,20]. Whilst this procedure enacted on the Q_A statistic furnishes attractive, weak instrument robust causal estimates, initial simulation results (not reported here) showed limited benefit of estimating δ in this way therefore we did not investigate potential implementation further.

There are a number of limitations to this work. The test statistic and weak instrument robust estimation requires an estimate of the covariance between the error in the estimated effect of each SNP on each exposure. Our simulation results highlight how important this data can be as the estimated values of F_{TS} and $\hat{\beta}_{Q,het}$ are changed so they become uninterpretable when this covariance is fixed to zero. Although this data is generally not available we propose a method to estimate it, using the phenotypic correlation between the exposures, which can be used to obtain a reasonable approximation if the relevant covariance when each SNP only explains a small proportion of each exposure. Where the data used to estimate the correlation between the exposures is the same data used to estimate the SNP-exposure associations the estimated value of $\Sigma_{V,j}$ will closely match the true value. When this is not the case the level of error in F_{TS} and $\beta_{Q,het}$ that results from misspecification of ρ_{km} will depend on the other parameters in the model. For F_{TS} this will depend on how related the exposures are and how strongly (or weakly) they are predicted. Misspecification of the conditional F-statistic will not matter if it is notably larger (or smaller) than 10 for all possible values of $\Sigma_{V,j}$ as this will not change the interpretation of the results. For $\beta_{Q,het}$ how much the specification of $\Sigma_{V,j}$ matters will depend on the estimated effect of each exposure, if all or all but one of these are zero $\Sigma_{V,j}$ will not affect the estimated results, and the magnitude of any effect will depend on the size of these estimated effects. When the data used for the estimation of ρ_{km}

does not match that used to obtain the SNP-exposure associations we have therefore proposed that the researcher investigates how variation in ρ_{km} affects the obtained values, and interpretation of F_{TS} and $\hat{\beta}_{Q,het}$. Where plausible variation in ρ_{km} does affect the interpretation of the results this limitation, and the resulting potential interpretations of the results obtained, should be accounted for by the researcher applying this method.

Another weakness of the test statistics provided here is the lack of standard errors for the point estimates of the direct effect of each exposure. We propose using a jackknife to estimate these standard errors. This does however make the estimation of these statistic more computationally intensive than would the case if the standard errors could be calculated analytically.

The weak instrument robust point estimates are robust to weak instruments but cannot produce reliable estimates when instruments become very weak or if only a small number of SNPs are available. Although we show this method works with moderately weak instruments it is not clear exactly how weak is too weak, or indeed how few instruments are too few, to produce either reliable point estimates or heterogeneity statistics. Gaining a more precise understanding of these questions is a topic for further research.

Although we propose weak instrument robust estimation, if the weak instruments are limited to only a small number of the exposures in the model an alternative approach may be to drop exposures (one at a time) until the conditional F-statistics show that all of the exposures are strongly predicted by the SNPs. This would however need to be considered carefully by the researcher. The model to be estimated should not be decided purely by which exposures can be predicted but driven by a research question of interest and dropping exposures has the potential to introduce directional pleiotropy into the estimation biasing the resulting effect estimates. The choice of approach to take would depend on the number of SNPs and exposures in the estimation and the relationship between the exposures as well as how weak the SNPs are as instruments. As illustrated by our application these approaches could be combined, excluding exposures until instrument strength is high enough to reasonably apply the weak instrument robust methods. The choice approach needs to be considered on a case by case basis.

Additionally although our final estimation $\beta_{Q,het}$ is robust to balanced pleiotropy it will still give biased estimates in the presence of unbalanced or directional pleiotropy. Multivariable MR Egger[14], has been proposed as a method for obtaining reliable MVMR estimates in the presence

of directional pleiotropy. Extending this approach to account for weak instrument bias is another topic of further research.

Box 1: Summary of statistics discussed in this paper.

Instrument strength statistics;

F - Measure of the strength of the instruments to predict one exposure. Applies to individual or summary level data and to univariable or multivariable MR estimation.

Conditional F-statistic F_{SW} - Measure of the strength of instruments to predict one exposure conditional on the other exposures included in the estimation. Applies to multivariable MR estimation with individual level data.

Conditional F-statistic F_{TS} - Measure of the strength of instruments to predict one exposure conditional on the other exposures included in the estimation. Applies to multivariable MR estimation with summary data.

Q_{x_j} - A Q-statistic from which F_{TS} is calculated.

Heterogeneity statistics;

Q_{IVW} - A heterogeneity test for MVMR that uses the IVW point estimates and does not account for the uncertainty in the estimated SNP-exposure associations. This test over rejects the null in the presence of weak instruments.

$Q_{IVW,up}$ - A heterogeneity test for MVMR that uses the IVW point estimates but accounts for the uncertainty in the estimated SNP-exposure associations. This test over rejects the null in the presence of weak instruments, but to a lesser extent that Q_{IVW}.

Q_A - A heterogeneity test for MVMR that is robust to weak instruments, in the sense that it has the appropriate type 1 error rate in the presence of weak instruments.

Estimation statistics;

$\hat{\beta}_{IVW}$ - Estimates of the causal effect of each exposure on the outcome, estimated using standard inverse variance weighting.

$\hat{\beta}_{Q}$ - Estimates of the causal effect of each exposure on the outcome, estimated through minimisation of Q_A. Robust to weak instruments.

$\hat{\beta}_{Q,het}$ - Estimates of the causal effect of each exposure on the outcome, estimated through minimisation of Q_A with an additional parameter to account for heterogeneity. Robust to weak instruments and pleiotropy.

Box 2: Recommended tests in Two-sample MVMR.

In all two-sample summary data MVMR estimation two statistics should be calculated;

1. Conditional F statistics, F_{TS}, for each exposure.

   These test the strength of the genetic variants to predict each exposure in the multivariable mode. $F_{TS} < 10$ suggests potential weak instrument bias in the MVMR estimation.

2. A Q-statistic for heterogeneity, Q_A, for the model.

   Rejection of Q_A using standard significant levels (e.g. $p < 0.05$) indicates potential pleiotropy in the form of excessive heterogeneity in the MVMR model. However, this test will often reject in the presence of weak instruments.

If weak instruments are detected, i.e. any of the F_{TS} values are less than 10, IVW MVMR estimates are potentially biased. When large numbers of SNPs are available this can be corrected through;

3. Estimating $\hat{\beta}_{Q,het}$ for each exposure

   This method gives estimates of the direct effect of each exposure on the outcome that are robust to (moderately) weak instruments.

4. An updated $Q_{A,min}$ which minimises the Q statistic over β_Q.

   This test provides a test for heterogeneity that has the correct size in the presence of weak instruments. Rejection of $Q_{A,min}$ using standard significant levels (e.g. $p < 0.05$) indicates potential pleiotropy in the MVMR model even in the presence of moderately weak instruments.

All of these tests and estimation statistics are provided in the MVMR R package.

Data availability

Details and data for the data used in this paper are available at https://github.com/WSpiller/MRChallenge2019. We additionally use some data from ALSPAC, details of the ALSPAC cohort and data access are available at http://www.bristol.ac.uk/alspac.

Code availability

The code used to conduct the simulations and applied analysis is available at https://github.com/eleanorsanderson/MVMRweakinstruments. The MVMR package is available at https://github.com/WSpiller/MVMR/.

Author contributions

ES and JB devised the project. ES conducted the analysis and wrote the first draft of the paper. WS developed the software package. All authors reviewed and approved the final version.

Funding

ES is funded through the MRC Integrative Epidemiology Unit (grant codes $MC\_UU\_00011/1$, $MC\_UU\_00011/2$). WS is supported by a Wellcome Trust studentship (108902/B/15/Z). JB is funded by an Expanding Excellence in England (E3) grant awarded to the Diabetes research group at the University of Exeter. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. The collection of the ALSPAC data used in this publication was funded by Wellcome (Grant ref: 093820/Z/19/Z). This publication is the work of the authors and they will serve as guarantors for the contents of this paper.

Acknowledgements

We are grateful for helpful discussions with Frank Windmeijer and Zoltan Kutalik during the development of this paper.

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

References

[1] George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

[2] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

[3] Stephen Burgess and Simon G Thompson. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology*, 181 (4):251–260, 2015.

[4] Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48(3):713–727, 2019.

[5] Doug Staiger and James Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

[6] James H Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. chapter 5 in identification and inference in econometric models: Essays in honor of thomas j. rothenberg, edited by dwk andrews and jh stock, 2005.

[7] Eleanor Sanderson and Frank Windmeijer. A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221, 2016.

[8] Debbie A Lawlor. Commentary: Two-sample mendelian randomization: opportunities and challenges. *International journal of epidemiology*, 45(3):908, 2016.

[9] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human molecular genetics*, 27(R2):R195–R208, 2018.

[10] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[11] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[12] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala Sheehan, and John Thompson. A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. *Statistics in medicine*, 36(11):1783–1802, 2017.

[13] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46(6):1985–1998, 2017.

[14] Jessica MB Rees, Angela M Wood, and Stephen Burgess. Extending the mr-egger method for multivariable mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in medicine*, 36(29):4705–4718, 2017.

[15] Andrew J Grant and Stephen Burgess. Pleiotropy robust methods for multivariable mendelian randomization. *arXiv preprint arXiv:2008.11997*, 2020.

[16] Johannes Kettunen, Ayse Demirkan, Peter Wurtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikainen, Matti Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. *Nature communications*, 7:11122, 2016.

[17] Stephen Burgess, Frank Dudbridge, and Simon G Thompson. Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906, 2016.

[18] Neil M Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in medicine*, 34(3):454–468, 2015.

[19] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, Qingyuan Zhao, Debbie A Lawlor, Nuala A Sheehan, John Thompson, and George Davey Smith. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International journal of epidemiology*, 12 2018. doi: 10.1093/ije/dyy258.

[20] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, Dylan S Small, et al. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48(3):1742–1769, 2020.

[21] Fabiola Del Greco M, Cosetta Minelli, Nuala A Sheehan, and John R Thompson. Detecting pleiotropy in mendelian randomisation studies with summary data and a continuous outcome. *Statistics in medicine*, 34(21):2926–2940, 2015.

[22] Paul Mitchell, Gerald Liew, Bamini Gopinath, and Tien Y Wong. Age-related macular degeneration. *The Lancet*, 392(10153):1147–1159, 2018.

[23] Katie L Pennington and Margaret M DeAngelis. Epidemiology of age-related macular degeneration (amd): associations with cardiovascular disease phenotypes and lipid factors. *Eye and vision*, 3(1):34, 2016.

[24] Lars G Fritsche, Wilmar Igl, Jessica N Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L Bragg-Gresham, Kathryn P Burdon, Scott J Hebbring, Cindy Wen, Mathias Gorski, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134, 2016.

[25] Stephen Burgess and George Davey Smith. Mendelian randomization implicates high-density lipoprotein cholesterol–associated mechanisms in etiology of age-related macular degeneration. *Ophthalmology*, 124(8):1165–1174, 2017.

[26] Qiao Fan, Joseph C Maranville, Lars Fritsche, Xueling Sim, Chui Ming Gemmy Cheung, Li Jia Chen, Mathias Gorski, Kenji Yamashiro, Jeeyun Ahn, Augustinus Laude, et al. Hdl-

cholesterol levels and risk of age-related macular degeneration: a multiethnic genetic study using mendelian randomization. *International journal of epidemiology*, 46(6):1891–1902, 2017.

[27] Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature Communications*, 11(1):1–11, 2020.

[28] Andy Boyd, Jean Golding, John Macleod, Debbie A Lawlor, Abigail Fraser, John Henderson, Lynn Molloy, Andy Ness, Susan Ring, and George Davey Smith. Cohort profile: the 'children of the 90s' -the index offspring of the avon longitudinal study of parents and children. *International journal of epidemiology*, 42(1):111–127, 2013.

[29] Abigail Fraser, Corrie Macdonald-Wallis, Kate Tilling, Andy Boyd, Jean Golding, George Davey Smith, John Henderson, John Macleod, Lynn Molloy, Andy Ness, et al. Cohort profile: the avon longitudinal study of parents and children: Alspac mothers cohort. *International journal of epidemiology*, 42(1):97–110, 2012.

[30] Neil Martin Davies, W David Hill, Emma L Anderson, Eleanor Sanderson, Ian J Deary, and George Davey Smith. Multivariable two-sample mendelian randomization estimates of the effects of intelligence and education on health. *Elife*, 8, 2019.

[31] Emma L Anderson, Laura D Howe, Kaitlin H Wade, Yoav Ben-Shlomo, W David Hill, Ian J Deary, Eleanor C Sanderson, Jie Zheng, Roxanna Korologou-Linden, Evie Stergiakouli, et al. Education, intelligence and alzheimers disease: evidence from a multivariable two-sample mendelian randomization study. *International journal of epidemiology*, 49(4):1163–1172, 2020.

[32] Eleanor Sanderson, George Davey Smith, Jack Bowden, and Marcus Munafo. Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. *Nature Communications*, 10, 2019.

[33] Tom G Richardson, Eleanor Sanderson, Tom M Palmer, Mika Ala-Korpela, Brian A Ference, George Davey Smith, and Michael V Holmes. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PLoS medicine*, 17(3):e1003062, 2020.

[34] Tom G Richardson, Eleanor Sanderson, Benjamin Elsworth, Kate Tilling, and George Davey Smith. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. *bmj*, 369, 2020.

[35] Alice R Carter, Eleanor Sanderson, Gemma Hammerton, Rebecca Richmond, George Davey Smith, Jon Heron, Amy Taylor, Neil M Davies, and Laura D Howe. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *BioRxiv*, page 835819, 2019.

[36] Alice R Carter, Dipender Gill, Neil M Davies, Amy E Taylor, Taavi Tillmann, Julien Vaucher, Robyn E Wootton, Marcus R Munafo, Gibran Hemani, Rainer Malik, Sudha Seshadri, Daniel Woo, Stephen Burgess, George Davey Smith, Michael V Holmes, Ioanna Tzoulaki, Laura D Howe, and Abbas Dehghan. Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. *BMJ*, 365, 2019.